Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: A privacy-protecting remote access system for health-related research and evaluation ☆

Kerina H. Jones *, David V. Ford, Chris Jones, Rohan Dsilva, Simon Thompson, Caroline J. Brooks, Martin L. Heaven, Daniel S. Thayer, Cynthia L. McNerney, Ronan A. Lyons

College of Medicine, ILS2, Swansea University, Swansea, Wales SA2 8PP, UK

ARTICLE INFO

Article history:
Received 26 July 2013
Accepted 7 January 2014
Available online xxxx

Keywords:
Data linkage
Remote access system
Privacy-protection
e-Records research

ABSTRACT

With the current expansion of data linkage research, the challenge is to find the balance between preserving the privacy of person-level data whilst making these data accessible for use to their full potential. We describe a privacy-protecting safe haven and secure remote access system, referred to as the Secure Anonymised Information Linkage (SAIL) Gateway. The Gateway provides data users with a familiar Windows interface and their usual toolsets to access approved anonymously-linked datasets for research and evaluation. We outline the principles and operating model of the Gateway, the features provided to users within the secure environment, and how we are approaching the challenges of making data safely accessible to increasing numbers of research users. The Gateway represents a powerful analytical environment and has been designed to be scalable and adaptable to meet the needs of the rapidly growing data linkage community.

1. Introduction

As many countries worldwide increasingly move towards electronically-held records across the range of public services, there is a growing potential for rich and novel studies through the record-level linkage of these administrative data. This is commonly referred to as data linkage, that is, merging two or more separate datasets containing information about the same individuals so that their topic areas can be studied together (e.g. hospital admissions and education). Indeed, many are of the view that we have a duty to re-use the wealth of data that are routinely collected as part of healthcare (and other public service) delivery to benefit patients and the public. However, there are competing challenges that have to be addressed, namely, how to ensure that individual privacy is protected whilst making the data as accessible as possible. Although some countries have linkable population-based health and social welfare registers [1], large-scale data linkage research is still a fairly novel area with relatively few long-established units, such as those in Australia [2], Canada [3,4], Scotland [5], England [6], as well as the Secure Anonymous Information Linkage System (SAIL) system in Wales [7]. However, it is an area that is developing rapidly with existing work being extended and new units being created. For example, the data linkage infrastructure is being extended across Australia [8], and four new Centres have been established in the UK: two in England, one in Scotland and one in Wales, creating the Farr Health Informatics Research Institute [9]. Among these is the Centre for the Improvement of Population Health through E-records Research (CIPHER) [10], based in Swansea (Wales), which is underpinned by the work of the SAIL system and aims to open up new opportunities to increase collaboration in data linkage research.

1.1. Motivation

The SAIL system was established in 2006 and, by using a range of technical and procedural privacy-protecting techniques, has brought together a wealth of health-related routinely-collected datasets in Wales, so that they can be reliably and anonymously linked at the individual level and securely used for research. The

Abbreviations: AIX, Advanced Interactive eXecutive; ALF, Anonymous Linking Field; CIPHER, Centre for the Improvement of Population Health through E-records Research; DB2, a family of database server products developed by International Business Machines (IBM); DP, Data Provider; HTTPS, HyperText Transfer Protocol Secure; IGRP, Information Governance Review Panel; LSOA, Lower Super Output Area; NHS, National Health Service; NWIS, NHS Wales Informatics Service; RALF, Residential Anonymous Linking Field; SAIL, Secure Anonymised Information Linkage; SQL, Structured Query Language; UKSeRP, UK Secure Research Platform; VPN, Virtual Private Network.

* Corresponding author. Fax: + 44 (0) 1792 513430.
E-mail address: k.h.jones@swansea.ac.uk (K.H. Jones).

1532-0464/$ - see front matter © 2014 The Aurthors. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2014.01.003

decision to establish the SAIL system as a repository with a central architecture, was made because of a number of factors. Among the major of these were that: many of the existing IT systems in health and social care settings were old and some were unstable; identity management, anonymisation and record-linkage would have been more problematic if attempting to operate in real time; there were unknown issues relating to data quality; there were varying degrees of organisational readiness and resources; and, it was more acceptable to Data Providers (DPs) to provide their anonymous data to SAIL, than for SAIL to plug into their systems and extract data as required. Further considerations were timeliness, computing capacity and cost [7].

There are various risks and attack models that need to be mitigated and controlled for in working with large-scale person-based data linkage in a repository model. It is essential that data are transported securely to avoid the risk of data loss or data falling into the hands of parties not authorised to receive them. A reliable and consistent matching technique is needful to ensure data accuracy, without which researchers could not have confidence in the data. Even though the data are anonymised, someone with legitimate access to the data, or a potential intruder, may attempt to re-identify individuals or clinicians. It is essential, therefore, that anonymisation is robust, that measures to further encrypt key variables are in place, and that data presented can be limited to the needs of a given project. As well as this, researchers need to understand their responsibilities and agree to abide by them. Furthermore, research outputs need to be scrutinised before dissemination to mitigate the risk of re-identification due to rare events or conditions. The database and analysis environment need to be secured against intruders and inappropriate data access. External scrutiny of the whole system is valuable for verification of compliance with Information Governance, for recommendations for improvement and to promote public and DP confidence.

Thus the SAIL objectives are: (1) ensuring data transportation is secure; (2) operating a reliable record matching technique to enable accurate record linkage across datasets; (3) anonymising and encrypting elements of the data to minimise the risk of re-identification; (4) applying measures to address disclosure risk in data views created for researchers; (5) ensuring data access is controlled and authorised; (6) scrutinising proposals for data utilisation and approving output; and (7) gaining external verification of compliance with Information Governance. These objectives still hold and their principles have been described previously [7,11], but there have been some important recent advances, the most notable being the way in which data are accessed. The SAIL model has always been to make data available to view, but not to remove; data are not passed outside of the system to researchers unless informed participant consent has been obtained. This is an important safeguard against linkage attack as it means researchers cannot take the data and link it with publicly available information to attempt to re-identify individuals. However, in the early years, approved researchers were only able to access data views via dedicated, secure, on-site terminals. This had practical disadvantages and limited the number of concurrent data users due to space constraints and travel requirements. The increasing numbers of researchers wishing to make use of SAIL data, from as far afield as Australia [12], made it impractical to continue with an on-site only data access model. This led to the development of the SAIL Gateway which is a remote data access system and analysis environment, as a powerful means of providing greater access to data without compromising individual privacy. This case study describes the main recent advances that have been made in the SAIL system for the curation, management and re-use of health-related data, with particular focus on the SAIL Gateway.

1.2. Paper organisation

The remainder of this paper is organised as follows: We briefly summarise the high-level architecture of the SAIL system for completeness, and to highlight recent developments against the SAIL objectives since the system was first described [7]. We then set out the principles on which the SAIL Gateway was developed, and use these to describe its operating model, the analysis environment, how file traffic in and out of the Gateway is managed, and plans for future scalability. We illustrate the user interface and set out the data user journey with the stages to be followed when engaging with SAIL. We discuss SAIL in context and describe how it can enable data linkage research to be carried out effectively and securely in a 'positive-sum' approach: that is, aiming to optimise the balance between data security and data utility [13]. As part of this, we incorporate discussion on the challenges we have encountered, and plans for improvement. We note that we sometimes use the terms 'researcher' and 'data user' interchangeably, whilst acknowledging that not all uses of the data are research.

1.3. Relevance to special issue on informatics methods in medical privacy

The special issue focuses on informatics methods in medical privacy: healthcare information collection and communication, healthcare data and knowledge management, healthcare information systems and technologies, and healthcare policies. Within these themes there are articles on research, practical applications, critical analyses and position papers. With its focus on practical applications for a privacy-preserving infrastructure and healthcare data management, we believe this article on the SAIL Gateway and its associated methodologies is highly relevant to the special issue.

2. Sail system architecture

The high-level architecture of the SAIL system is illustrated in Fig. 1. SAIL uses formal data sharing agreements with DPs and supports them in their due diligence processes to provide datasets to SAIL in accordance with Information Governance. The SAIL technical team provides guidance to the DP on the file extract specification and this uses a split-file approach to enact the separation principle. The defining feature of this principle is that the commonly-recognised identifiers, and other fields used in the matching process, (consisting of name, address, postal code, gender and date of birth (designated File 1)) are separated from the clinical or event-based descriptive data (such as disease codes and prescriptions (designated File 2)). A distinction is made between the two types of data to reflect the varying potential for direct (File 1) or indirect (File 2) attribution of variables to an identity, and so that the required variables can be used in the matching process, without the associated descriptive data being present. File 1 is sent to the NHS Wales Informatics Service (NWIS) [14] which acts as a Trusted Third Party (TTP), and File 2 is sent directly to SAIL. There are two methods of data transportation: file upload using a secure electronic transfer facility (known as the NHS switching service); or if the DP has a secure file download service, SAIL and NWIS can make use of this. In the past SAIL also used portable data transfer media, such as CDs, but we no longer allow this for improved security.

NWIS carries out matching and anonymisation, whereby the commonly-recognised identifiers are replaced with an Anonymous Linking Field (ALF) assigned uniquely to each person represented in the File 1 dataset, along with minimal demographics to create File 3. Matching is carried out against the Welsh Demographic Service, an administrative register of people in Wales registered with

Please cite this article in press as: Jones KH et al. A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: A privacy-protecting remote access system for health-related research and evaluation. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.01.003

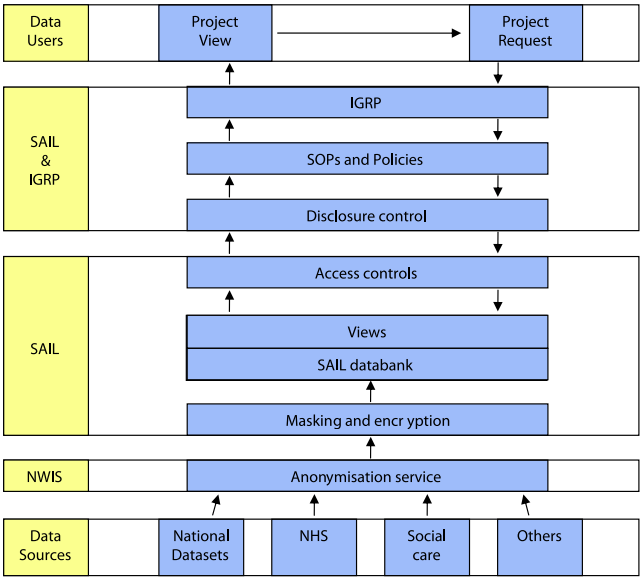

Fig. 1. SAIL architecture. This diagram shows the SAIL databank system and the controls in place for data acquisition and utilisation, with an indication of the roles carried out by each party. Beginning at the base of the diagram, SAIL has formal agreements with data providers to provide their data to the databank in accordance with Information Governance. The commonly-recognised identifiers are anonymised at NWIS, who provide a trusted third party service to SAIL. Further processes of masking and encryption are carried out at SAIL, and the SAIL databank is constructed. From the top of the diagram, requests to use the data are reviewed by SAIL and an independent Information Governance Review Panel (IGRP) to assess compliance with Information Governance before access can be allowed. Once this is agreed, a data view is created by SAIL staff, and access to this view can be made available via the SAIL Gateway. For this to happen, further data transformations are carried out to control the risk of disclosure, and the data user signs an access agreement for responsible data utilisation, in accordance the specifications of the IGRP to comply with Information Governance.

a General Practitioner, and it acts as a proxy for a Welsh demographic database [7,11]. The ALF is a unique number, based on the person's National Health Service (NHS) number, encrypted using a Blowfish algorithm [15]. The minimal demographics consist of week of birth (in place of date of birth), a simple code for gender, and Lower Super Output Area (LSOA, approximately 1500 population) of residence in place of postal code. File 3 is then sent to SAIL for recombination with File 2 to create an anonymous version of the original dataset [7]. The minimal demographics and their granularities were decided upon based on the balance of data security and data utility. Having these variables associated with the descriptive data within SAIL means that important issues such as gender, age, birth seasonality, and area of residence can be studied in relation to health, without revealing identity.

Advances in work with geographic data have enabled a parallel process to be developed for address-level datasets. In this way, a unique Residential Anonymous Linking Field (RALF) is assigned to each address in the File 1R dataset, similarly creating File 3R which can be recombined with the corresponding event-based dataset (File 2) at SAIL. An example of this sort of dataset is local authority housing characteristics. Through the technologies that have been developed we are able to associate a RALF with the ALFs within it, so that the health of individuals in relation to their living environment can be studied. However, great care must be taken so that individual privacy is not compromised. This is because of the possibility of being able to work out the location of a given property in certain scenarios. For example, the density of housing in a rural area is often considerably lower than in an urban setting. This could possibly lead to re-identification of individuals living within a given house and, therefore, it is essential to have robust methods to ensure privacy is protected when working with residence-based data [16,17].

Further processes of masking and encryption are carried out at SAIL, for example: the ALF is encrypted to create an ALF-E (Encrypted), to ensure that no one accessing the data (whether at SAIL or at NWIS) can decrypt the field; and codes pertaining to individual clinics or General Practices are masked to respect professional identities. Subject to quality assurance, the dataset is incorporated into the SAIL databank. The SAIL databank operates on a DB2 platform (Data Warehouse Edition on Advanced Interactive eXecutive (AIX)) running on an IBM 'P' series computer: DeepBlue-C.

SAIL has created a management system, referred to as SAIL Info Central. This is used by SAIL staff to manage dataset information and researcher access. It also has a data user interface, and this is described in Section 4. Requests to use the data are initially reviewed by members of the SAIL team to assess feasibility. This entails assessing whether the requested data and variables are available, and whether the work can be completed within the proposed timescale and resource allocation. It is then reviewed by an independent Information Governance Review Panel (IGRP) to assess compliance with Information Governance before data access can be agreed [7]. Some DPs request the right to review proposals seeking to use their data, and SAIL complies with this requirement. The IGRP is comprised of members of professional and regulatory bodies, and we have recently increased the lay representation, such that it now includes two members of the SAIL Consumer Panel, which was established in 2011 [18]. The IGRP provides feedback on the appropriateness of the proposal to the public interest, and highlights any perceived risks in linking the requested data. This information is used by the SAIL team to guide data view preparation. As well as disclosure measures applied at a databank level (such as masking of practitioner codes) SAIL uses a variety of measures for risk mitigation in data views created for researchers. These can be tailored to the needs of the specific project and include: aggregation and suppression; limiting numbers of variables provided/sequential provision; and project-specific encryption of the ALF-E to prevent cross-linkage where data users are involved in multiple projects. This is carried out with dialogue between the researcher and a senior analyst to ensure that data utility and security are optimised. The data view is then made available within a specified schema, beyond which the researcher has no access. However, with the increase in numbers of people wishing to use SAIL data, this can be an onerous task and we are looking to develop a more automated process.

Once approval is granted, a data view is created, and access to this view can be provided, subject to the researcher signing a data access agreement for responsible data utilisation and understanding that misconduct will be subject to disciplinary action. Data access can only occur via the SAIL Gateway and this is discussed in more detail below.

3. SAIL Gateway principles

The SAIL Gateway has been created on four basic principles to ensure that it is able to meet the needs of the growing data linkage community [8,9], and these are to:

(A) Operate a remote access system that provides secure data access to approved users.
(B) Host an environment that provides a powerful platform for data analysis activities.
(C) Have a robust mechanism for the safe transfer of approved files in and out of the system.
(D) Ensure that the system is efficient and scalable to accommodate a growing data user base.

Please cite this article in press as: Jones KH et al. A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: A privacy-protecting remote access system for health-related research and evaluation. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.01.003

4. SAIL Gateway case description

4.1. SAIL Gateway operating model

The operating model of the SAIL Gateway, that is how it operates to accomplish its function, is represented in Fig. 2. SAIL data are stored on DB2 nodes with controls at various levels of the model to ensure that only approved personnel are able to access the data for specified purposes. Only highly-trusted, on-site technical staff have direct access to the raw data, and this is for data curation and management purposes. Even so, no one at SAIL has access to identifiable data due to the File 1/File 2 separation principle and anonymisation of File 1 at NWIS. All staff accessing data, for whatever purpose, are required to abide by the data access agreement for responsible data use. The Gateway has a range of security measures, and these include firewalls, two-factor authentication methods, encrypted network connections, and security servers. Approved data users are able to access their data view remotely via their own computer. To be able do this, data users are provided with an account to log onto the Gateway via a Virtual Private Network (VPN) and a YubiKey authentication token which, when placed in the USB slot, conveys a one-time password as if it was entered via the keyboard [19]. The user is then able to access a provisioned remote desktop as the only route to access the data view. The Gateway system uses Active Directory group policies to control the configuration of the remote desktop, and users are prevented from copying or transferring the data, or mapping network drives by these functions being disabled. Further safeguards are provided by user-level logging of all SQL commands issued (including by off-site users). This is enabled via the audit feature on DB2 so that all activity is recorded and tracked, including logins (successful and attempted), all queries sent to the database, and all objects (tables or views) accessed by the data user. This is useful for monitoring system efficiency and also to track suspected instances of misconduct.

4.2. SAIL Gateway analysis environment

There are many features that comprise the SAIL Gateway analysis environment. The user interface has been designed to provide data users with a familiar Windows environment and to host an array of toolsets and applications so that it forms a flexible research platform. A version of SAIL Info Central is made available externally to the SAIL Gateway but subject to having a Gateway account (Fig. 3), and this provides data users with information about datasets, support options and the services available to them. The main features available within the SAIL Gateway are illustrated in Fig. 4. There is a WIKI [20] which acts as a central point of information on dataset dictionaries, training materials, best practice guidelines, tips on data querying, and FAQs. It also signposts data users to an instant messaging system [21] for rapid communication between logged-in analysts and to a discussion forum [22] for Questions and Answers. Data users have access to an NHS Clinical Terminology Browser which enables them to search for read codes and their meanings to assist them in formulating SQL queries. SQL scripts are developed in an IBM InfoSphere environment, and querying tools and commonly used applications, such as MS Office, SPSS, R and STATA, are included as standard; but users can also request the addition of specific applications for which they hold a licence. Computationally heavy tasks are usually carried out via SQL querying of DB2 on DeepBlue-C, whereas less demanding requirements can be met by the applications running on the remote desktop. As well as this, data users can request permission to import non-data files, such as syntax scripts and reference documents to support their analysis and meet their particular needs. Data users also have access to SAIL Info Central inside the Gateway.

4.3. SAIL Gateway file transfers

Even though the data within SAIL are anonymous, the standard operating model is that row-level data are not permitted to leave SAIL, except when all relevant regulatory and governance approvals, including appropriate participant informed consent, have been obtained. This principle is an acknowledgement of the potential risk of disclosure that may be posed by the release, particularly, of multi-variable data. When a data user has completed their analysis they are not able to remove their results from the Gateway, as this can only be done by a SAIL data guardian. The role of the guardian is carried out by a trusted senior analyst who manually scrutinises the proposed outputs to ensure that the risk of disclosure has been mitigated. For example, in line with Office of National Statistics guidance for administrative data, no outputs should contain row-level data or table cell counts less than 5, but

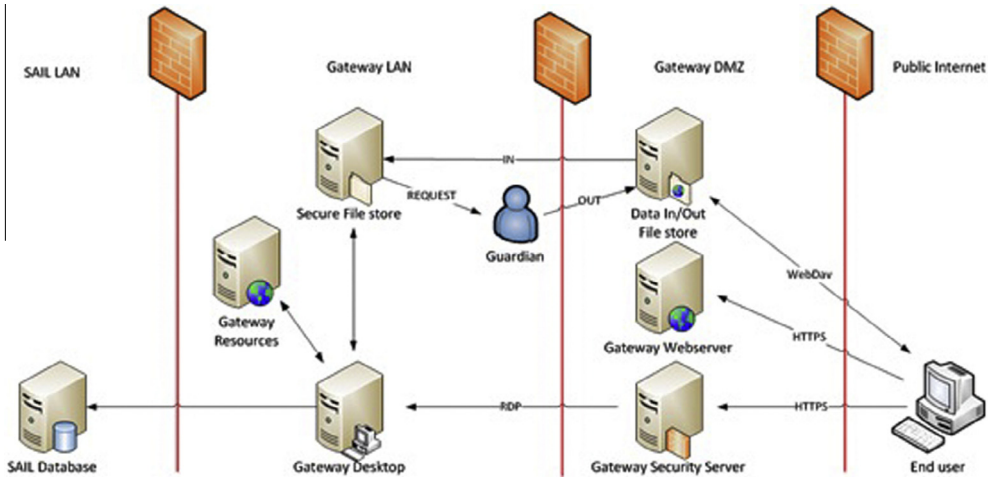

Fig. 2. The SAIL Gateway. The SAIL Gateway is a remote data access system, and this simplified illustration shows how data are accessed by end users using their own computer. Once approved with an account, data users are provided with a Gateway desktop within the Gateway Local Area Network (LAN), and this is accessed via remote desktop protocol (RDP) interposed by a Gateway security server. The Gateway desktop communicates with the SAIL database in the SAIL LAN to provide data users with a specified data view. Within the Gateway LAN, data users have access to analysis tools, a secure file store and other resources. File transfers into and out of the Gateway LAN are mediated by a guardian.

Please cite this article in press as: Jones KH et al. A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: A privacy-protecting remote access system for health-related research and evaluation. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.01.003

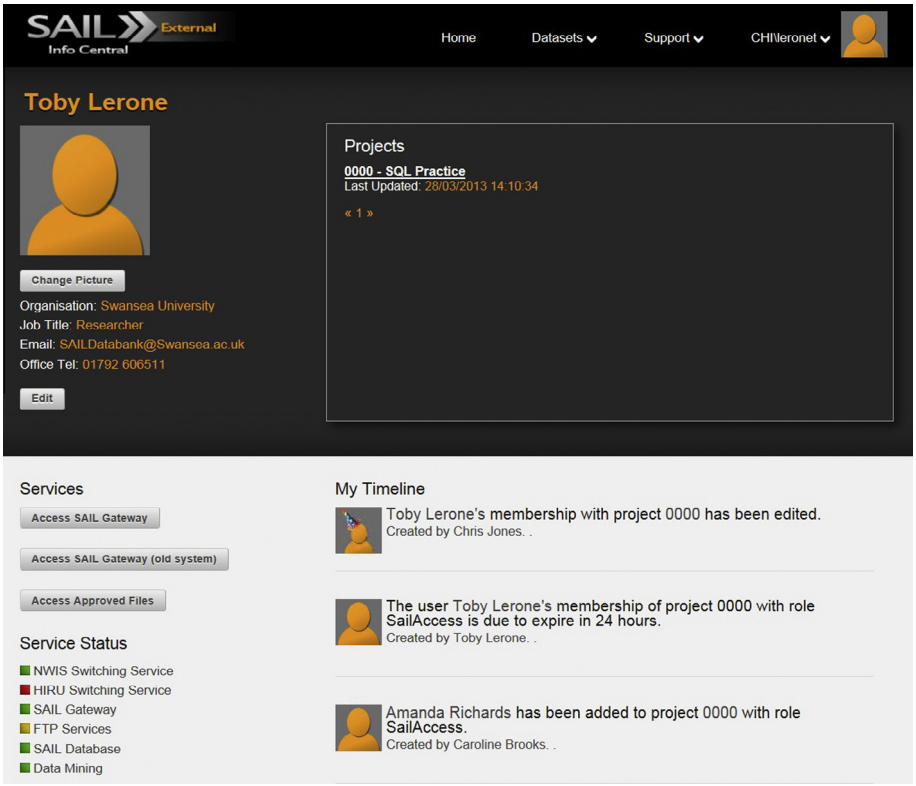

Fig. 3. SAIL Info Central screenshot. This screenshot displays the home page for a data user on the SAIL Info Central (External to the SAIL Gateway) site. The top section displays menus to more information about the datasets and support options. The top left hand section displays information about the user, which is editable by the user. The bottom left hand section displays the services available to the user and indicates service status. The centre section displays all the projects that the user is authorised to access and the hyperlink directs them to more information about the project. The bottom right hand section display a timeline of news feed from projects and dataset updates.

this may need to be higher for particularly sensitive data [23]. This process can be labour-intensive and we are developing more automated processes to reduce demand on staff time. As well as using core SAIL datasets, data users are also able to request that additional datasets, such as those collected in a research study, are uploaded to the system and anonymously linked to existing data within SAIL. These are subject to our standard quality assurance procedures scrutiny to ensure privacy is protected [7].

4.4. SAIL Gateway efficiency and scalability

SAIL benefits from high performance computing infrastructure to store and manage the data, currently comprising over 4 billion rows relating to approximately 4 million historical and current residents of Wales. This requires a substantial commitment to infrastructure and its management, but it ensures there are no constraints in storage capacity and efficiency of data processing. A recent in-house efficiency audit showed that the average execution time for SQL queries constructed by data users is less than 30 s. To date over 100 projects have been approved to use SAIL data with further applications in progress. These have included clinical trials, disease registers, cohorts, and observational, epidemiological and methodological studies. Some examples are illustrated briefly here. The SAFER (Support and Assessment for Fall Emergency Referrals) studies are a series of trials designed to improve assessment and referral of people who fall. They use routine data in conjunction with data collected in the studies [24]. The UK Multiple Sclerosis (MS) Register has an innovative model combining data from NHS Neurology clinics, routine sources and directly from people with MS in the form of Patient Reported Outcome Measures [25–28]. A similar model has been used by the Ankylosing Spondylitis cohort study [29]. The Housing Regeneration and Health study is using local authority housing data with SAIL data to assess whether improvements in social housing lead to tangible health benefits [30]. Various other projects make use of the array of SAIL datasets without incorporating additional data. Among these are a study on the risk of gastro-intestinal infections in connection with the use of proton-pump inhibitors [31], and others on methods to enhance selection of trial participants [32,33]. None of these projects would have been practicable without a data linkage infrastructure.

There are currently 75 remote desktops and, due to increasing demand, we aim to increase this to support over 300 remote desktops in the next 2 years, but more could be supported if required. This will require us to continually review and improve the efficiency of our processes. Although SAIL was primarily established to hold Welsh datasets, the principles are applicable to datasets from other sources for researchers wishing to use the Gateway environment as a Safe Haven and analysis platform. Specific adaptations can be made to meet other specifications, such as the secure data lab approach for particularly sensitive data. This entails setting up a 'safe room' with stringent and strictly controlled data access conditions, and is a method used by organisations such as ONS and the UK Data Service [21,34].

A flowchart illustrating the data user's journey is shown in Fig. 5. A prospective data user makes contact with the SAIL team and sets out their proposal for IGRP review. Following approval and the creation of the user account and data view, the data user is able to conduct their analysis within the SAIL Gateway environment. Results are scrutinised before release of the approved outputs for dissemination.

Information on the management structures in place for the SAIL system, providing data to SAIL, using SAIL data and summaries of the research portfolio and outputs, can be found on the SAIL

Please cite this article in press as: Jones KH et al. A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: A privacy-protecting remote access system for health-related research and evaluation. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.01.003

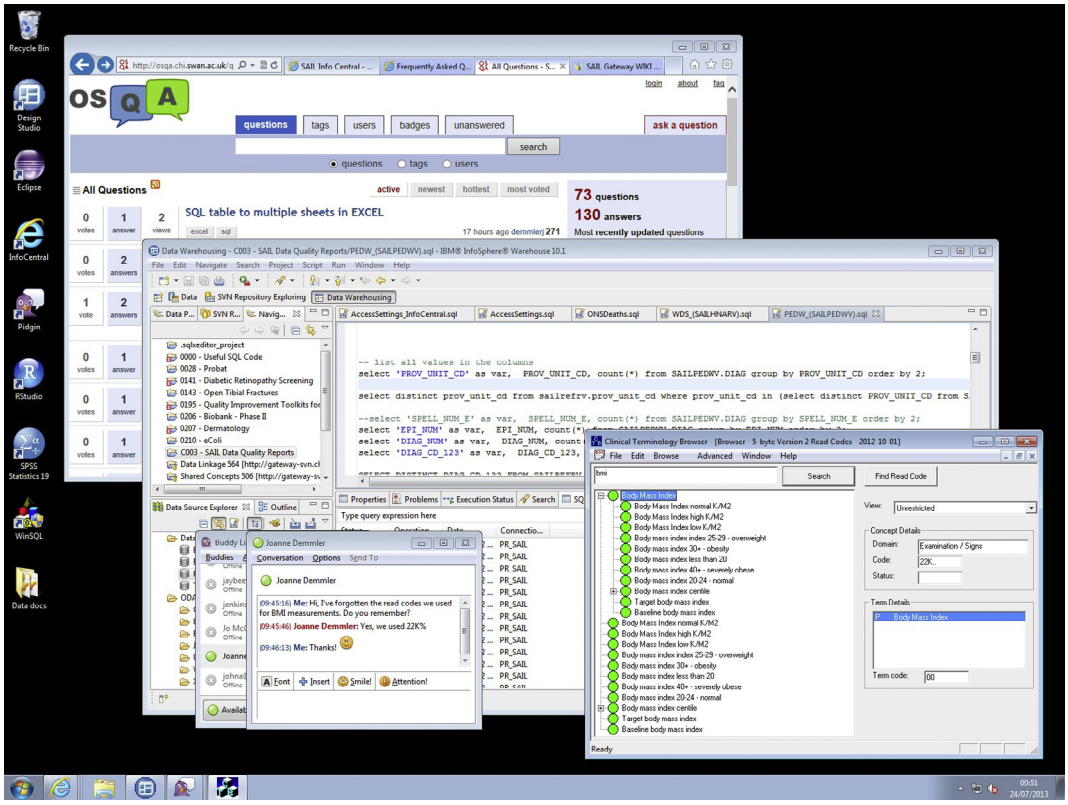

Fig. 4. SAIL Gateway screenshot. This screenshot displays the main features available within the SAIL Gateway. The left hand side displays the desktop icons to software that is installed as default. Starting from the top and working clockwise the screenshot displays the internet based resources – Question and Answer Forum, Frequently Asked Questions, SAIL Info Central (Internal to the SAIL Gateway), and the SAIL Gateway WIKI. The second screen displays the IBM InfoSphere environment used to develop SQL code to manipulate SAIL data to be research ready. The third screen displays the NHS Clinical Terminology Browser which enables users to search for read codes and their meanings. Finally the last two screens display the instant messaging system that enables users to communicate and share information within the secure SAIL Gateway environment.

website [35]. As well as internal management, the SAIL system is subject to a rolling programme of independent audit. This is valuable in such a demanding field as it provides us with recommendations for improvement, and then the assurance provided by the audit reports can be conveyed to stakeholders. We are currently working towards ISO 27001 compliance to further increase robustness and stakeholder confidence. This is an extensive process requiring considerable time and effort from the SAIL team, but it is a valuable exercise to ensure that we operate to the highest standards.

The main principles of SAIL methodologies are summarised in Table 1 and the importance of recent developments, the most notable changes being due to the development of the SAIL Gateway, are highlighted.

5. Discussion and conclusion

The SAIL system is a national architecture for e-health research and evaluation, and it has been designed to operate on a positive-sum approach [13] whereby anonymised person-level data can be made available for research and individual privacy can be maintained. The use of such data presents more benefits, but also more risks, than aggregated data [36]. We have used a range of approaches in establishing the SAIL system and there have been multiple challenges to address, many of which are on-going. These have included DP confidence in data provision, the development of technical and procedural methods for data management and access, having an appropriate technical and management infrastructure, differences in quality and coding systems between datasets, staff training and capacity, and navigating the Governance frameworks. There are other ways in which the challenges could have been addressed, but the SAIL system has been developed through a pragmatic approach, to create a timely solution that combines technical processes, control measures, authorisations and accountability. As a result, the SAIL system is able to operate with the support of multiple DPs and in compliance with Information Governance. Furthermore, the remote access afforded by the SAIL Gateway provides convenience to data users to promote research collaboration within a secure safe haven and powerful analysis environment. Other architectures may be possible in the future, and we are working with collaborators to explore these in seeking continual improvement.

The data linkage landscape is evolving rapidly, and the systems that exist and those in development have a variety of operating models. SAIL both compares and contrasts with aspects of various other models, and some key examples are given here. Among the earlier developments in data linkage research was the establishment of the Oxford Medical Record Linkage System (Ox-Link). Through this was created the Oxford Record Linkage Study, comprising 10 million records pertaining to 5 million people in England between the 1960s and the 1990s. In common with SAIL, Ox-Link was based on a repository model and it used a NHS Central Register to provide a reliable matching standard. However, the Ox-Link policy was to comprehensively link all the records, rather than to prepare them on an *ad hoc* basis, and this contrasts with SAIL where linkage between datasets is not made on a wholesale basis [6]. The Western Australia data linkage system was established in the 1990s and this uses a different operating model to that of SAIL. It

Please cite this article in press as: Jones KH et al. A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: A privacy-protecting remote access system for health-related research and evaluation. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.01.003

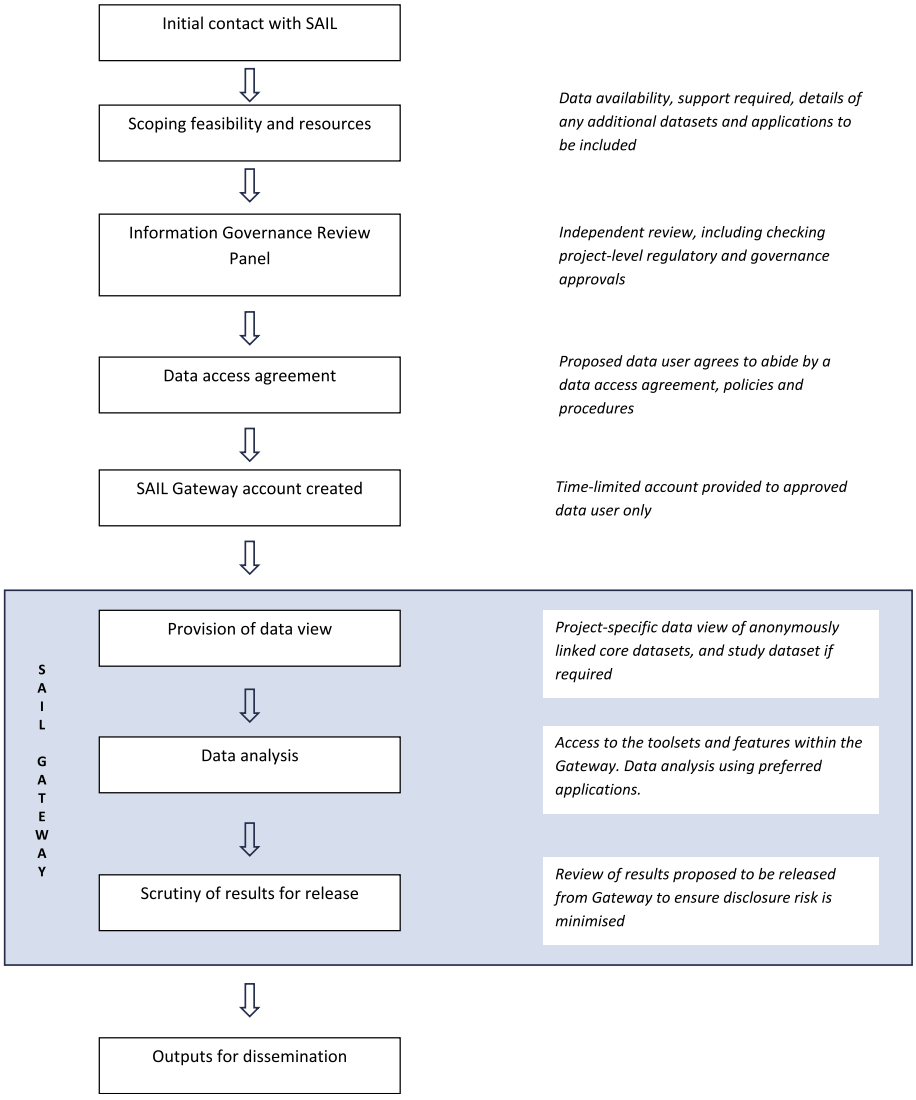

Fig. 5. The SAIL Gateway data user journey. This flowchart illustrates the SAIL data user journey from initial contact with SAIL to dissemination of outputs. Work conducted within the SAIL Gateway is highlighted.

is not based around a data repository, but instead the datasets remain with the DPs and there is a database of master linkage keys to enable records from different DPs to be linked together. As a further contrast, the data and project-specific linkage keys are provided externally to the researcher, whereas in SAIL, analysis is carried out within the SAIL Gateway environment [2]. Further developments are underway to extend and enhance the data linkage infrastructure across Australia [8].

Manitoba Health houses multiple datasets with an anonymised population-based research registry playing a central role, and linkages are made when required for particular studies, similar to in SAIL. However, unlike SAIL which uses a TTP, Manitoba Health receives the commonly-recognised identifiable information (name, address, etc.), analogous to the SAIL File 1, to carry out a matching process and assign an anonymous identifier to each individual-level record. But separation is maintained and DP permission is required before the content of their administrative datasets (analogous to SAIL File 2) are provided to be linked and used for research [3]. The standard operating model of SAIL is similar to that of Population Data BC in that data access is provided remotely and so similar safeguards apply, such as the ability to suspend data access if malpractice is suspected. However, there are some key differences, for example, SAIL uses a TTP and engages in research whereas Population Data BC itself operates as a TTP for data linkage and foregoes a research function [4,37]. Data linkage is well established in Scotland and developments are underway on the further enhancement of a flexible data linkage infrastructure in connection with the Scottish element of the Farr Institute for Health Informatics Research [9]. This is incorporating an indexing service, linkage agent, safe haven facilities and a one-stop to assist researchers [5,38]. Clearly, there are already key areas of overlap with the SAIL system, and inter-operability across the four Farr Centres, within a proportionate Governance framework, is a key priority.

There are a variety of effective models across the data linkage landscape, and although they differ in operational aspects, their common remit is to make anonymous linked administrative data available for use by approved third parties for legitimate uses within the relevant regulatory and governance frameworks. There are many challenges to be addressed in meeting the needs of high quality research whilst complying with Information Governance and, as shown, these can be approached in different ways. The over-arching governance framework for SAIL has been designed to cover the full data life-cycle: from pre-data acquisition negotiations with data providers to the release of results for publication. However, in such a rapidly expanding field, there will be more

Please cite this article in press as: Jones KH et al. A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: A privacy-protecting remote access system for health-related research and evaluation. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.01.003

Table 1
The SAIL system objectives, methods and recent developments.

	Objective	Methods	Developments
1	Ensuring data transportation is secure	Following data provider Information Governance permissions and subject to a data sharing agreement, datasets are split into a demographic component (comprising the commonly-recognised identifiers), and a clinical or event component (such as medication records and procedures). These are transported to NWIS [13] and to SAIL, respectively, using HTTPS (web based secure file upload via switching service, or secure file download)	The use of secure file transfers is more robust than using portable transfer media such as CDs or USBs, and less subject to data mis-direction or loss
2	Operating a reliable record matching technique to enable accurate record linkage across datasets	Matching and assignment of a consistent, unique ALF to each individual is carried out by NWIS acting as a TTP so that SAIL does not handle identifiable data. An ALF can be assigned to NHS and non-NHS datasets (such as local authority housing, or fire service datasets). Similarly, a RALF has been developed for residences in address-level datasets [14,15]	The extension of the matching and record-linkage processes beyond healthcare data, and to include RALFs, opens up new dimensions for research
3	Anonymising and encrypting the data to minimise the risk of re-identification	Demographic data are anonymised and encrypted by the TTP and subjected to quality assurance to ensure content anonymity. SAIL receives only the ALF, week of birth, gender code and area of residence (LSOA), which are then recombined with the clinical/event component of the dataset. Further encryption of the ALF is carried out at SAIL to form the ALF-E. A parallel process is in place for the RALF. Linkage across datasets is made via the ALF-E and RALF	The use of the ALF and RALF means that datasets can be linked at the individual record and address level (respectively), enabling a wide range of research whilst preserving privacy
4	Applying measures to address disclosure risk in data views created for researchers	A variety of measures can be applied at creation of individual data views to maximise utility and minimise disclosure risk, including: masking of practitioner codes; aggregation and suppression; limiting numbers of variables provided/ sequential provision; project-specific encryption of the ALF-E to prevent cross-linkage where data users are involved in multiple projects	Having a variety of measures ensures a flexible approach. However, the current method can be labour-intensive for a senior analyst as the number of data users grows and more automated methods are in development
5	Ensuring data access is controlled and authorised	Subject to data user verification, a data access agreement, and physical and procedural controls, data users are assigned a time-limited account to access their data view. Whereas previously this was location-specific, data users are now able to access data remotely via the SAIL Gateway. Safeguards include a fire-walled VPN, enhanced user authentication, the use of YubiKeys, logging of all SQL commands, and configuration controls to ensure that data cannot be removed or transferred unless authorised. Datasets and user access are managed via SAIL Info Central	The SAIL Gateway enables greater numbers of researchers to engage safely with SAIL, compared to the previous access model, which was on-site only. The Gateway environment provides users with a range of familiar tools and applications, with secure file transfers in and out of the Gateway. SAIL Info Central provides a centralised management system and user interface
6	Scrutinising proposals for data utilisation and approving output	All proposals to use SAIL data are subject to review by an independent IGRP. In addition, some DPs request the right to review proposals seeking to use their data, and SAIL complies with this requirement. Output is scrutinised for potential disclosure risk before results can be released	The lay representation on the IGRP has been increased to enhance the patient/public viewpoint. Additional automation in privacy-protecting measures is being developed and will further streamline the process of output scrutiny
7	Gaining external verification of compliance with Information Governance	As well as in-house monitoring, IG compliance is verified by a regular programme of independent audit	SAIL is also working towards ISO 270001 compliance

The objectives of the SAIL system are shown, along with a brief summary of the methods in place and recent developments.

challenges to be surfaced and addressed. For example, we have identified the need for more automated privacy-protecting methods in data view preparation and results release. This is needful for more efficient workload management as demand for data increases, and for creating more quantifiable risk mitigation measures. However, it is not a trivial task as the data requirements (datasets, variables and granularities) vary depending on the research questions to be addressed in a particular study, so that one size does not fit all. We also aim to extend the SAIL Gateway to form a UK Secure Research Platform (UK SeRP). The aim of UK SeRP is to act as a Safe Haven and analysis platform for any *bone fide* researcher who requires a secure data linkage environment to work on datasets they legitimately hold. This is being taken forward through our work in CIPHER [10], and will introduce new stakeholders, DPs and requirements. As the governance and data linkage landscape develops we will aim to continually review and improve the SAIL system, in collaboration with data linkage partners, to aim to stay at the cutting edge of data linkage research for the benefit of the population.

Contributorship statement

All authors made substantial contributions to the conception and design of this work, and the acquisition of data. KHJ drafted the article and all authors were involved in revising it critically for important intellectual content. All authors gave final approval of the version to be published.

Funding

This work was supported by the National Institute of Social Care and Health Research (NISCHR), Wales, with Grant No. ISS3.

Acknowledgments

We would like to acknowledge the support of NHS Wales Informatics Service (our Trusted Third Party), the many data providers who supply datasets to SAIL, our Consumer Panel and our many

collaborators and partners who work with us on research studies, governance and other aspects of the SAIL system.

References

[1] Gissler M, Haukka J. Finnish health and social welfare registers in epidemiological research. Nor Epidemiol 2004;14(1):113–20. <http://www.ntnu.no/ojs/index.php/norepid/article/view/284/262>.
[2] Kelman CW, Bass AJ, Holman CDJ. Research use of linked health data – a best practice protocol. ANZ J Public Health 2002;26:251–5.
[3] Roos LL, Brownwell M, Lix L, Roos NP, Walld R, Macwilliam L. From health research to social research: privacy, methods, approaches. Soc Sci Med 2008;66:117–29.
[4] Hertzman CP, Meagher N, McGrail KN. Privacy by Design at Population Data BC: a case study describing the technical, administrative, and physical controls for privacy-sensitive secondary use of personal information for research in the public interest. J Am Med Inform Assoc 2013;20:1 25–28. http://dx.doi.org/10.1136/amiajnl-2012-001011 [Published Online First: 30.08.12].
[5] Scottish Government. A blueprint for health records research in Scotland. July 2012. <http://www.scot-ship.ac.uk/sites/default/files/Reports/SHIP_BLUEPRINT_DOCUMENT_final_100712.pdf> [accessed 18.05.13].
[6] Gill LE. OX-LINK: the Oxford medical record linkage system. In: Record linkage techniques. Oxford: University of Oxford; 1997. p. 19.
[7] Ford DV, Jones KH, Verplancke J-P, Lyons RA, John G, Brown G, et al. The SAIL Databank: building a national architecture for e-health research and evaluation. BMC Health Serv Res 2009;9:157.
[8] Boyd JH, Ferrante AM, O'Keefe CM, Bass AJ, Randall SM, Semmens JB. Data linkage infrastructure for cross-jurisdictional health-related research in Australia. BMC Health Serv Res 2012;12:480. http://dx.doi.org/10.1186/1472-6963-12-480.
[9] Medical Research Council (b); 2012. <http://www.mrc.ac.uk/Newspublications/News/MRC008799> [accessed 18.05.13].
[10] Centre for the Improvement of Population Health through Ehealth Research. <http://www.swan.ac.uk/medicine/cipher/> [accessed 18.05.13].
[11] Lyons RA, Jones KH, John G, Brooks CJ, Verplancke J-P, Ford DV, et al. The SAIL databank: linking multiple health and social care datasets. BMC Med Inform Decis Mak 2009;9:3.
[12] Gabbe BJ, Lyons RA, Lecky FE, Bouamra O, Woodford M, Coats TJ, et al. Comparison of mortality following hospitalisation for isolated head injury in England and Wales, and Victoria, Australia. PLoS One 2011;6:e20545 [Published 31.05.11].
[13] Cavoukian A. Privacy by design... take the challenge. Information and Privacy Commissioner of Ontario; 2009. <http://www.ipc.on.ca/english/Resources/Discussion-Papers/Discussion-Papers-Summary/?id=856> [accessed 18.05.13].
[14] NHS Wales Informatics Service. <http://www.wales.nhs.uk/sitesplus/956/home> [accessed 18.05.13].
[15] Schneier B. Description of a new variable-length key, 64-bit block Cipher (Blowfish). Fast Software Encryption, Cambridge security workshop proceedings (December 1993). Springer-Verlag; 1994.
[16] Rodgers SE, Lyons RA, Dsilva R, Jones KH, Brooks CJ, Ford DV, et al. Residential Anonymous Linking Fields (RALFs): a novel information infrastructure to study the interaction between the environment and individuals' health. J Public Health 2009:1–7.
[17] Rodgers SE, Demmler JC, D'Silva R, Lyons RA. Protecting health data privacy while using residence-based environment and demographic data. Health Place 2011. http://dx.doi.org/10.1016/j.healthplace.2011.09.006.
[18] Jones KH, McNerney CL, Ford DV. Involving consumers in the work of a data linkage research unit. Int J Consum Stud 2014;38(1):45–51. http://dx.doi.org/10.1111/ijcs.12062.
[19] YubiKey Standard. <http://www.yubico.com/products/yubikey-hardware/yubikey/> [accessed 23.07.13].
[20] Screwturn wiki. <http://www.screwturn.eu/> [accessed 18.05.13].
[21] Pidgin: universal chat client. <http://www.pidgin.im/> [accessed 18.05.13].
[22] Osqa: the open source Q&A system. <http://www.osqa.net/> [accessed 18.05.13].
[23] Office of National Statistics. <http://www.ons.gov.uk/ons/guide-method/best-practice/disclosure-control-policy-for-tables/index.html> [accessed 18.05.13].
[24] Snooks HA, Anthony R, Chatters R, Cheung WY, Dale J, Donohoe R, et al. Support and assessment for fall emergency referrals (SAFER 2) research protocol: cluster randomised trial of the clinical and cost effectiveness of new protocols for emergency ambulance paramedics to assess and refer to appropriate community based care. BMJ Open 2012;2:e002169. http://dx.doi.org/10.1136/bmjopen-2012-002169.
[25] Ford DV, Jones KH, Middleton RM, Lockhart-Jones H, Maramba IDC, Noble GJ, et al. The feasibility of collecting information from people with Multiple Sclerosis for the UK MS register via a web portal: characterising a cohort of people with MS. BMC Med Inform Decis Mak 2012;12(1):73 [18.07.12].
[26] Jones KH, Ford DV, Jones PA, John A, Middleton RM, Lockhart-Jones H, et al. A large-scale study of anxiety and depression in people with multiple sclerosis: a survey via the web portal of the UK MS register. PLoS One 2012;7(7):e41910. http://dx.doi.org/10.1371/journal.pone.0041910.
[27] Jones KH, Ford DV, Jones PA, John A, Middleton RM, Lockhart-Jones H, et al. The physical and psychological impact of Multiple Sclerosis using the MSIS-29 via the UK MS register. PLoS One 2013 [Published 31.01.13].
[28] Jones KH, Ford DV, Jones PA, John A, Middleton RM, Lockhart-Jones H, et al. How people with Multiple Sclerosis rate their quality of life: an EQ-5D survey via the UK MS register. PLoS One 2013;8(6):e65640. http://dx.doi.org/10.1371/journal.pone.0065640.
[29] Atkinson MD, Brophy S, Siebert S, Gravenor MB, Phillips CJ, Ford DV, et al. Protocol for a population-based Ankylosing Spondylitis (PAS) cohort in Wales. BMC Musculoskelet Disord 2010;11:197. http://dx.doi.org/10.1186/1471-2474-11-197.
[30] Rodgers SE, Heaven M, Lacey A, Macey S, Poortinga W, Dunstan FD, et al. Cohort profile: the housing regeneration and health study. Int J Epidemiol 2012. <http://www.ncbi.nlm.nih.gov/pubmed/23179304>.
[31] Brophy S, Jones KH, Rahman MA, Zhou S-M, John A, Atkinson MD, et al. Incidence of Campylobacter and Salmonella infections following first prescription for PPI – a cohort study using routine data. Am J Gastroenterol 2013;108(7):1094–100. http://dx.doi.org/10.1038/ajg.2013.30.
[32] Brooks CJ, Stephens JW, Price DE, Ford DV, Lyons RA, Prior SL, et al. Use of a patient linked data warehouse to facilitate diabetes trial recruitment from primary care. Prim Care Diab, 2009. doi: http://dx.doi.org/10.1016/j.pcd.2009.06.004.
[33] Mcgregor J, Brooks C, Chalasani P, Chukwuma J, Hutchings H, Lyons RA, et al. The Health Informatics Trial Enhancement Project (HITE): using routinely collected primary care data to identify potential participants for a depression trial. Trials 2010;11:39. http://dx.doi.org/10.1186/1745-6215-11-39. <http://www.trialsjournal.com/content/11/1/39>.
[34] UK Data Service. <http://discover.ukdataservice.ac.uk/Catalogue/?sn=7196&type=Data%20catalogue> [accessed 30.11.13].
[35] SAIL databank. <http://www.saildatabank.com/> [accessed 23.07.13].
[36] El Emam K, Cavoukian A. A positive-sum paradigm in action in the health sector. Information and Privacy Commissioner of Ontario; 2010. <http://www.ipc.on.ca/images/Resources/positive-sum-khalid.pdf> [accessed 18.05.13].
[37] Population data BC. <http://www.popdata.bc.ca/aboutus> [accessed 18.05.13].
[38] Information services division. <http://www.isdscotland.org/Products-and-Services/eDRIS/> [accessed 02.12.13].

Please cite this article in press as: Jones KH et al. A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: A privacy-protecting remote access system for health-related research and evaluation. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.01.003